NPS-OR-99-003

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Determining Characteristic Groups to Predict Army Attrition

by

Samuel E. Buttrey
Harold J. Larson

February 1999

Prepared for: HQDA, ODCSPER
300 Army Pentagon
ATTN: DAPE-PRS RM 2C744
Washington, DC 20310-0300

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93943-5000

RADM Robert C. Chaplin
Superintendent

Richard Elster
Provost

This report was prepared for and funded by HQDA, ODCSPER, 300 Army Pentagon, Washington, DC 20310-0300.

This report was prepared by:

SAMUEL E. BUTTREY
Assistant Professor of
Operations Research

Harold J Larson
HAROLD J. LARSON
Professor of Operations Research

Reviewed by:

GERALD G. BROWN
Associate Chairman for Research
Department of Operations Research

Richard E. Rosenthal
RICHARD E. ROSENTHAL
Chairman
Department of Operations Research

Released by:

DAVID W. NETZER
Associate Provost and Dean of Research

REPORT DOCUMENTATION PAGE

Form approved

OMB No 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE February 1999	3. REPORT TYPE AND DATES COVERED Technical

4. TITLE AND SUBTITLE

Determining Characteristic Groups to Predict Army Attrition

5. FUNDING

MIPR 7MNPSO1002

6. AUTHOR(S)

Samuel E. Buttrey and Harold J. Larson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Postgraduate School
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

NPS-OR-99-003

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

HQDA, ODCSPER
300 Army Pentagon
ATTN: DAPE-PRS RM 2C744
Washington, DC 20310-0300

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The Office of the Deputy Chief of Staff, Personnel (ODCSPER), is charged with managing the Army's military strength levels and forecasting future strength levels for planning purposes. ODCSPER is reformulating its Enlisted Loss Inventory Model (ELIM), which projects losses of first-term enlisted personnel. These projections in turn are passed to a program which is designed to maintain the Army's strength as closely as possible to prescribed levels. These projections are based on characteristic groups, a set of sub-groups of recruits who are similar in terms of sex, education level, term of service and mental category; the presumption has been that attrition rates ought to be different between groups. However in recent years ELIM projections have been unsatisfactory.

This study used Classification and Regression Tree methodology (CART) to generate improved c-groups for predicting not only first-term attrition but also early-term behavior and re-enlistment. The most important variables by which to create these groups turn out to be race and gender. Generally white women have the lowest term completion and re-enlistment rates; those for non-white women and white men are similar; and those for non-white men are the highest.

14. SUBJECT TERMS

manpower, attrition, CART

15. NUMBER OF PAGES

28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL

Abstract

The Office of the Deputy Chief of Staff, Personnel (ODCSPER), is charged with managing the Army's military strength levels and forecasting future strength levels for planning purposes. ODCSPER is reformulating its Enlisted Loss Inventory Model (ELIM), which projects losses of first-term enlisted personnel. These projections in turn are passed to a program which is designed to maintain the Army's strength as closely as possible to prescribed levels. These projections are based on characteristic groups, a set of sub-groups of recruits who are similar in terms of sex, education level, term of service and mental category; the presumption has been that attrition rates ought to be different between groups. However in recent years ELIM projections have been unsatisfactory.

This study used Classification and Regression Tree methodology (CART) to generate improved c-groups for predicting not only first-term attrition but also early-term behavior and re-enlistment. The most important variables by which to create these groups turn out to be race and gender. Generally white women have the lowest term completion and re-enlistment rates; those for non-white women and white men are similar; and those for non-white men are the highest.

Executive Summary

The Office of the Deputy Chief of Staff, Personnel (ODCSPER), is charged with managing the Army's military strength levels and forecasting future strength levels for planning purposes. These forecasts have a great impact on the Army's Military Personnel Account (MPA); this single item represents about one third of the Army's budget and is the largest single account in the Department of Defense.

ODCSPER has a set of computer programs that make up its military strength management system. This redesigned system will use current hardware and software technology, and employ more recent, and easier to use, design concepts.

One piece of the system is the the Enlisted Loss Inventory Model (ELIM), which projects losses of first-term enlisted personnel. These ELIM projections are used in turn by an optimization model (COMPLIP, Computation Of Manpower Programs using LInear Programming) which is designed to maintain the Army's strength as closely as possible to prescribed levels.

This current ELIM model bases its projections on characteristic groups (c-groups), whose structure has remained unchanged since the strength management system was initially implemented. These c-groups partition first-term enlisted personnel according to sex, education level, mental category (AFQT group) and term of service in a specific way. In recent years, forecasts made by the ELIM model have not been satisfactory; this has caused ODCSPER to consider alternative partitions of first-term enlisted personnel.

This study used Classification and Regression Tree methodology (CART) to generate improved c-groups; these new c-groups are designed to differ in first-term retention rates, to the maximum extent possible. For this task the CART technique used only information about whether the recruit did or did not complete his or her first term. A separate analysis tried to distinguish three outcomes: non-completion of the first term, completion without re-enlistment, and both completion and re-enlistment. Finally, this method was also used to characterize differences in early-term attrition.

The CART methodology was able to define new c-groups that outperform the old in terms of misclassification rate. Addtionally, they have the desirable property of being less dissimilar in size than the current c-groups. Most interesting is the set of attributes that

the method selects to define the groups. The most important distinction turns out to be gender: the rate of term completion for women is about 45% in the particular data set we employed, while that for men is about 60%. (The actual numbers vary a bit from year to year, but the difference always exists.) Among women, race matters; white women have a higher attrition rate than non-whites. There is a disparity between white and non-white men that is somewhat smaller. Additional "splits" between groups are made on variables like length of term of service, AFQT scores, and level of education.

The same general trend holds true when considering re-enlistment together with retention. White women have the lowest re-enlistment rates; non-white women are similar to white men; and non-white men have the highest rates.The various college bonus programs are associated with lower re-enlistment rates, as we might expect, but interestingly the college bonus does not have much effect on term completion. (This may be partly because the bonus money is, in many cases, completely paid before the term is completed.) Another important variable is the Career Management Field, which describes generally the recruit's function in the Army.

Finally, the same race and gender distinctions hold true when considering attrition early in the first term. Graphs of attrition rate against month show that every group has a peak of attrition in the first few months of the first term; the rate then drops down to a steady state in about month 9. The peak is highest for white women and lowest for non-white males. The steady-state attrition rate, after month 9, is again highest for white women; for the other three the rates are essentially equal and constant over the remainder of the first term.

One reason that women in general, and especially white women, may have higher attrition rates is that their terms of service tend to be longer. Of course this is not relevant to the observation that the early-term rates are high. Still, white women have longer term lengths, on average, than other groups, for reasons that are not clear, and non-white men the shortest. Gender differences may be partly due to contract rates that differ by CMF since some CMFs are not open to women. However, the reasons for any racial differences are not known. In any case, the race and gender differences are real. Although a larger proportion of women than men sign up for four-year terms, for example, attrition rates are highest for white women when controlling for length of term.

1 Background

The Office of the Deputy Chief of Staff, Personnel (ODCSPER), is charged with managing the Army's military strength levels and forecasting future strength levels for planning purposes. These forecasts have a great impact on the Army's Military Personnel Account (MPA); this single item represents about one third of the Army's budget and is the largest single account in the Department of Defense.

For more than 20 years the Army has employed essentially the same military strength management system. It is used for modeling near-term needs for, and adjustments to, manpower levels, as well as for longer-term projections. This system contains a suite of individual computer programs which are loosely integrated overall and built on outdated hardware and software platforms. Routine modeling chores are time-consuming to perform, and the system is difficult and expensive to maintain or enhance. Long training periods are required to familiarize analysts with its complex user controls. ODCSPER is currently sponsoring a new design for its military strength management system to overcome these difficulties. The redesigned system will use current hardware and software technology, employing more recent, and easier to use, design concepts.

Much of the structure of the current military strength management system will be retained in the new system. In particular, the Enlisted Loss Inventory Model (ELIM) for projecting losses of first-term enlisted personnel will be maintained. These ELIM projections are used in turn by an optimization model (COMPLIP, Computation Of Manpower Programs using LInear Programming) which is designed to maintain the Army's strength as closely as possible to prescribed levels. This structure will also be retained.

This current ELIM model bases its projections on characteristic groups (c-groups), whose structure has remained unchanged since the strength management system was initially implemented. These c-groups partition first-term enlisted personnel according to sex, education level, mental category (AFQT group) and term of service in a specific way. They were originally designed, in part, to identify differences in first-term retention behavior, which in turn was expected to increase accuracy in short- and long-term forecasting accuracy. In recent years, forecasts made by the ELIM model have not been satisfactory; this has caused ODCSPER to consider alternative partitions of first-term enlisted personnel.

This study used Classification and Regression Tree methodology (CART) to generate c-groups for use with ODCSPER's new Military Strength Management System; these new c-groups are designed to differ in first-term retention rates, to the maximum extent possible. As this project continued, interest was also expressed in categorizing differences in retention in the early months of a recruit's first term. In addition, interest arose in groupings which distinguished three groups: those who did not complete the first term, those who did complete the first term but did not re-enlist, and those who did choose to re-enlist at the completion of the first term. CART has also been used for these efforts. The CART methodology is briefly described in the following section.

2 Description of CART

The basic approach used in CART was developed in Breiman *et al.* (1983) and in earlier papers by these authors. To briefly describe this methodology, assume we have a group of 10000 first-term enlistees and, for each, we have the following information:

a. Sex — Male or Female

b. Race — Three groups: W, N, O

c. Age — Calendar age at start of term, in months

d. Education — Grouped into 5 categories: GED,<HS,HS,C2,>C2

e. Term — 1 if term of service completed, 0 if not

Any number of additional variables could equally well be accommodated, but to keep the discussion simple we shall refer only to those listed above. Our goal is to partition the 10000 enlistees into a number of groups; the groups have no members in common and together account for all 10000 enlistees. We want the proportions of persons to complete their term of service (within a group) to differ as much as possible from one group to another.

CART attacks this problem in the following way. Assume that 5800 of the enlistees do in fact complete their term of service. Thus, in the full group, 58% complete their term of service. We want to describe the categorical variable Term (either term is completed or not) using the values of the other available variables: Sex, Race, Age, Education. Of these, Sex, Race and Education are categorical and Age is "continuous" or numeric. This is done by CART using a classification tree, since Term is taken to be categorical (not continuous). For classification trees, CART repeatedly splits any given group according to the value of the group's deviance; for the full group of 10000 enlistees being discussed this deviance is $-2\ln\left((.58)^{5800}(.42)^{4200}\right) = 13605.84$. This classification tree deviance is actually -2 times the log of the likelihood function, appropriate if the 10000 enlistees behave like independent Bernoulli trials with constant probability .58 of re-enlisting.

If the initial group of 10000 is split into any two groups of sizes n_1, n_2 (where $n_1 + n_2 = 10000$), we can likewise compute the deviance for each of the two groups. The sum of these two deviances can easily be shown to be no larger than the deviance for the original group of 10000 (and generally it will be smaller). In building its classification tree, CART considers each of the available variables (Sex, Race, Age, Education) in turn; for the categorical variables CART effectively partitions the levels into 2 sets (in all possible ways), uses each of these to split the enlistees into two groups and for each evaluates the resulting sum of the two group deviances. For continuous variables, CART partitions the range of the variable into two parts (between all possible successive values of the variable), again splits the enlistees accordingly and evaluates the sum of the two group deviances. That variable, and the split on that variable which produces the smallest sum of the two group deviances, is then used to partition the "root node" (original collection of all 10000 enlistees) into two "child nodes".

The same process is then applied to each of these two child nodes; *i. e.*, each child is now a parent node and all variables are again examined to find that variable, and the split of that variable, which results in the smallest possible sum of the deviances for the two child nodes. This then is used to split each of the two original child nodes into two new child nodes, now giving 4 nodes, and the process is again repeated. This process could conceivably be continued until each node is "pure" (every person in the node acts the same way with respect to completion of term and the node deviance is 0). The usual computer implementations of CART continue the node splitting until the number of cases in the node is "small" (say 10 or fewer) or until the difference between the parent deviance and the sum of the child deviances is "small" (say the drop in deviance is less than 1% of the parent deviance). The results of this procedure can be pictured as a classification tree, like Figure 1. The root node (at the top) contains the full collection of enlistees; the variable and value for that variable which causes the biggest drop in deviance is identified on the lines dropping from the root to two child nodes, and so on. When the process is stopped, the final nodes (rectangles) are called the leaves and the number of these leaves is called the size of the tree. The percentage written in each node is the proportion of enlistees in the node that completed the first term; the number written below each node is the number of enlistees contained in the node.

Thus, as seen in Figure 1, the first (most important) split is on sex, with 60.5% of males and only 44.9% of females completing their first term. The white females are then split off from the other races, giving two leaves describing female behavior. The males are first split on education, then on age and finally (for the better educated, older males) split on the race variable, giving a six leaf tree.

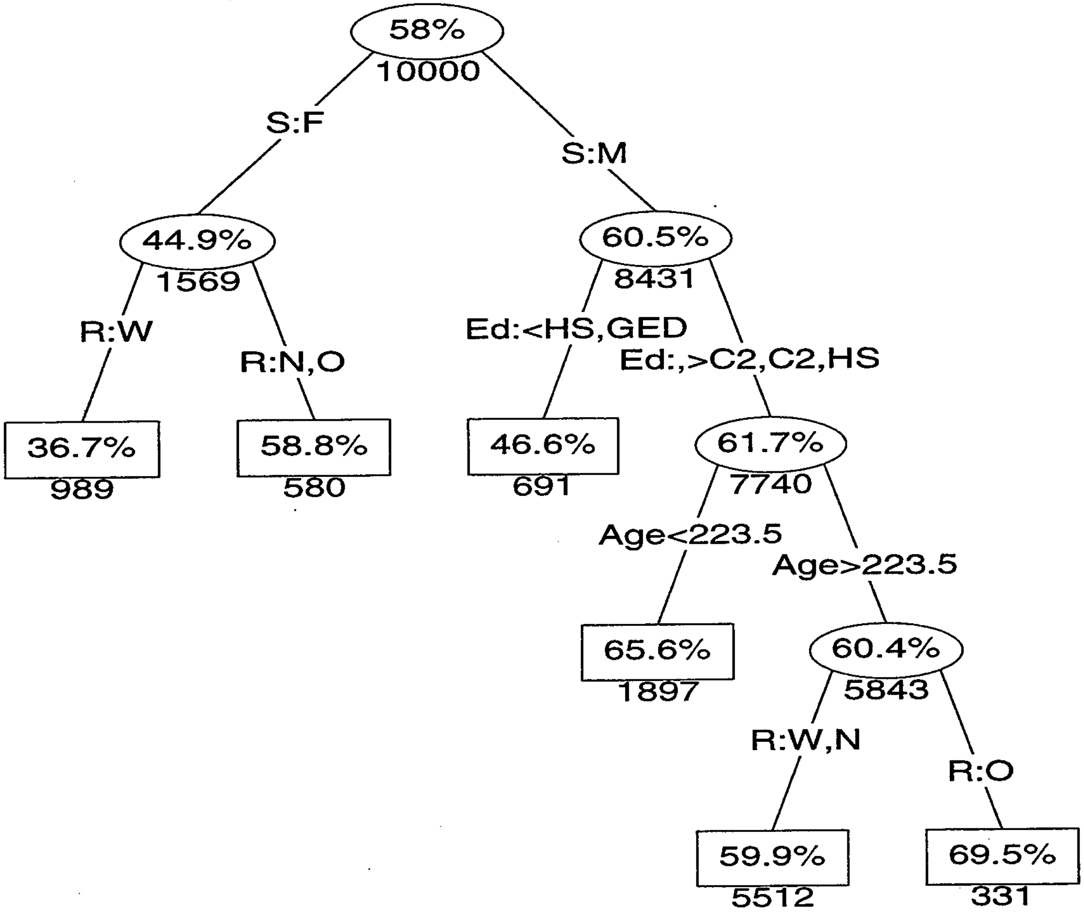
Different trees are frequently compared using their misclassification rate; for a classification tree, this computation assumes that everyone in the same node is classified the same, determined by the classification with the greatest frequency in the node. Thus, for the case of modeling completion of term, each node is classified in two ways: re-enlist (if more than 50% choose this option) or don't re-enlist (if less than 50% choose to re-enlist). For the root node at the top, the misclassification rate is .42 (the proportion not re-enlisting). The tree misclassification rate is given by summing the numbers misclassified across all leaves and dividing this total by 10000, the number of enlistees. Thus, for the tree in Figure 1, this numerator is

$$.367(989)+.412(580)+.466(691)+.344(1897)+.401(5512)+.305(331)=3887,$$

so the tree misclassification rate is .3887, about 92% of the misclassification rate at the root.

It is clear that none of the leaves in Figure 1 are pure. Why does this tree have only six leaves? The suggested way to determine the "best" size for a tree is to use the procedure called cross-validation. The default usage of this procedure is to take the original data set (10000 enlistees) and randomly split it into ten equal-size groups (1000 per group). Each group of 1000 is held out, in turn, and the remaining 9000 are used to build trees of each

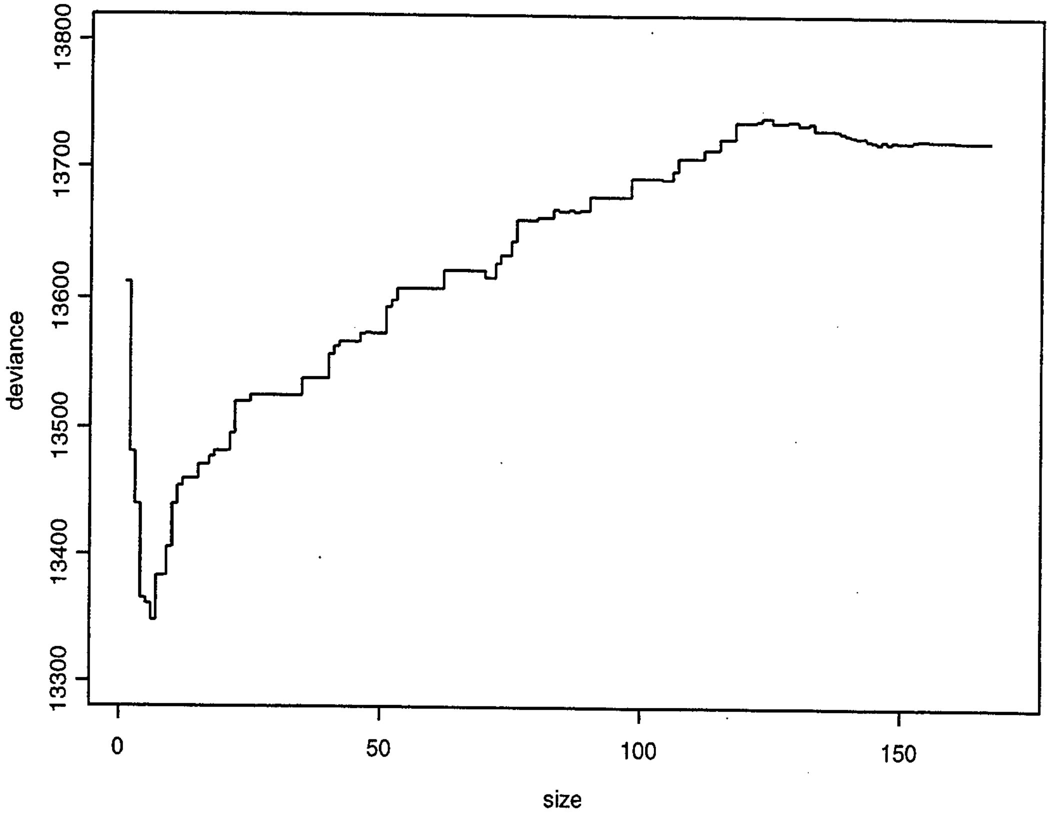
Figure 1: Classification tree for 10000 first-term enlistees.



Example of classification tree. Within each node is the proportion of recruits in that node who re-enlisted; beneath each node is the number of recruits in that node. Recruits are broken into groups based on Sex (S), Race (R), Education level (ED), and Age in Months (Age). For example, of 989 white females (leftmost node), 36.7% re-enlisted.

possible size (from 2 to roughly 170 leaves). The 1000 cases held back are dropped down each tree, which then determines a classification for each case, and the resulting deviance is computed. Thus for trees of size 2, 3, 4, ... we have 10 deviances and can find the minimum deviance of the 10, for each size. Then one can evaluate the penalized deviance, a weighted sum of the minimum deviance and the number of leaves in the tree, and plot the result, as in Figure 2. This has the typical shape of going through a minimum (at six

Figure 2: Cross-validation plot for tree in Figure 1.



Example of cross-validation. For each size of tree (that is, for each number of terminal nodes) there is an estimate of the deviance for trees of that size, penalized by the size. The "best" size is the one for which the penalized deviance is smallest. In this case, the "best" tree has six leaves.

leaves for these data) and then increasing. Since the minimum occurs at six leaves, we have some evidence that this may be the "best" size to use. This procedure is intended to avoid over-fitting the data, using a larger size tree than may be called for.

3 Data available

The initial data provided for this study were taken from the small tracking file, an

ODCSPER contractor-maintained data base. All first-term enlistees, from calendar year 1989 through June 30, 1987, were included in the data file. This file contains one line per enlistee, for this period, and includes social security number for identification purposes. The following six variables were included for each enlistee:

1. AFQT score (numeric, range 1 through 99),
2. race (4 codes),
3. sex,
4. term of service (2 through 6 years, plus missing),
5. civilian education code (33 different levels), and
6. age in months at start of term.

Each enlistee record also included "trailers", identifying various personnel transactions (completed term, immediate re-enlistment, reason discharged from service, among others) and the month in which the transaction occurred. The number of these trailer records varied from 1 (marking the person as nonprior service in month 1) and could go as high as 13. The file, and the trailer records in particular, were current through June 30, 1997. Thus, if a person joined the service in May, 1990, with a 6 year term, it is possible from this file to determine whether (s)he completed the first term of service. It is not possible to know with certainty whether a person who joined the service in January, 1996, with a two year term, completed the first term of service using this file.

Initial efforts concentrated on building classification trees to model completion of term, using the six variables listed above. While the trees did reduce the misclassification rate and the deviance, discussions during the first presentation of the work performed lead to a number of suggestions and clarifications. Primary among these was the fact that much of the first term attrition occurs during the early months after basic training; this phenomenon has been studied and is described later. Discussion also centered on the possibility that other demographic variables beyond the six listed above might prove useful in describing first term attrition; additionally, interest was expressed in using CART to model re-enlistments for first-term enlistees, in addition to modeling completion of term. These topics will also be discussed.

The small tracking file contains only the demographic variables listed earlier. Thus, the addition of more variables for use with CART required new sources of data, which could be melded with the existing records using the social security numbers. The Office of Economic and Manpower Analysis (OEMA) provided two useful files for this study. The first of these provided the ZIP code for home of record of the first-term enlistees, their Military Occupational Specialty (MOS) codes, and the length of time spent in DEP (delayed entry program) for all persons joining the Army during calendar years 1991 through 1997. The home of record (ZIP code) is frequently missing for enlistment dates prior to 1994. Additionally, OEMA provided information on losses to first term enlistees which were recorded between July, 1997 and May, 1998, to allow more up-to-date determination of completion of first term of service.

The United States Army Recruiting Command (USAREC) also provided data on non-prior service enlistees who joined during calendar years 1991 through 1997. These data included social security numbers, ages, as well as enlistment bonus and Army college fund information. These data were requested for two specific purposes. The first of these was to provide missing ages of enlistees (especially for the 1994 cohort) which were not available in the small tracking file. The second purpose was to investigate whether enrollment in the Army college fund program would prove useful in CART modeling of completion of first term and/or re-enlistment.

For use as variables in employing CART, the ZIP codes were melded into 6 regional groups: Northeast, Southeast, Midwest, Mountain, Far West and Other (Alaska, Hawaii, Puerto Rico). The MOS labels were combined into twelve Career Management Fields, using information provided by an ODCSPER contractor, modified by PERSCOM's Enlisted MOS Structure Chart dated 7/31/98. The USAREC data was used to construct a college fund variable with 4 levels (no college fund, two-year college fund, three-year college fund, four-year college fund) for use with CART. This information was also used to fill in missing entries for term of service; that is, if an enlistee had a missing term of service in the small tracking file, and (s)he was listed as participating in the three-year college fund, then the missing term of service was replaced with a three-year term. No existing terms of service (from the small tracking file) were changed using this college fund variable, although there are a few instances in which the two do not agree.

4 CART description of completion of term

As mentioned earlier, the original goal of this research was to use CART to model the completion of term for non-prior service enlistees. This requires knowledge of whether or not particular individuals completed the first term of service; since this first term of service can be anywhere from two to six years, it is clear that the data needed must span several years of observation. Table 1 gives a description of the available enlistee data for fiscal years 1991 through 1996, using trailer records from the small tracking file, as well as the OEMA data on enlistee losses.

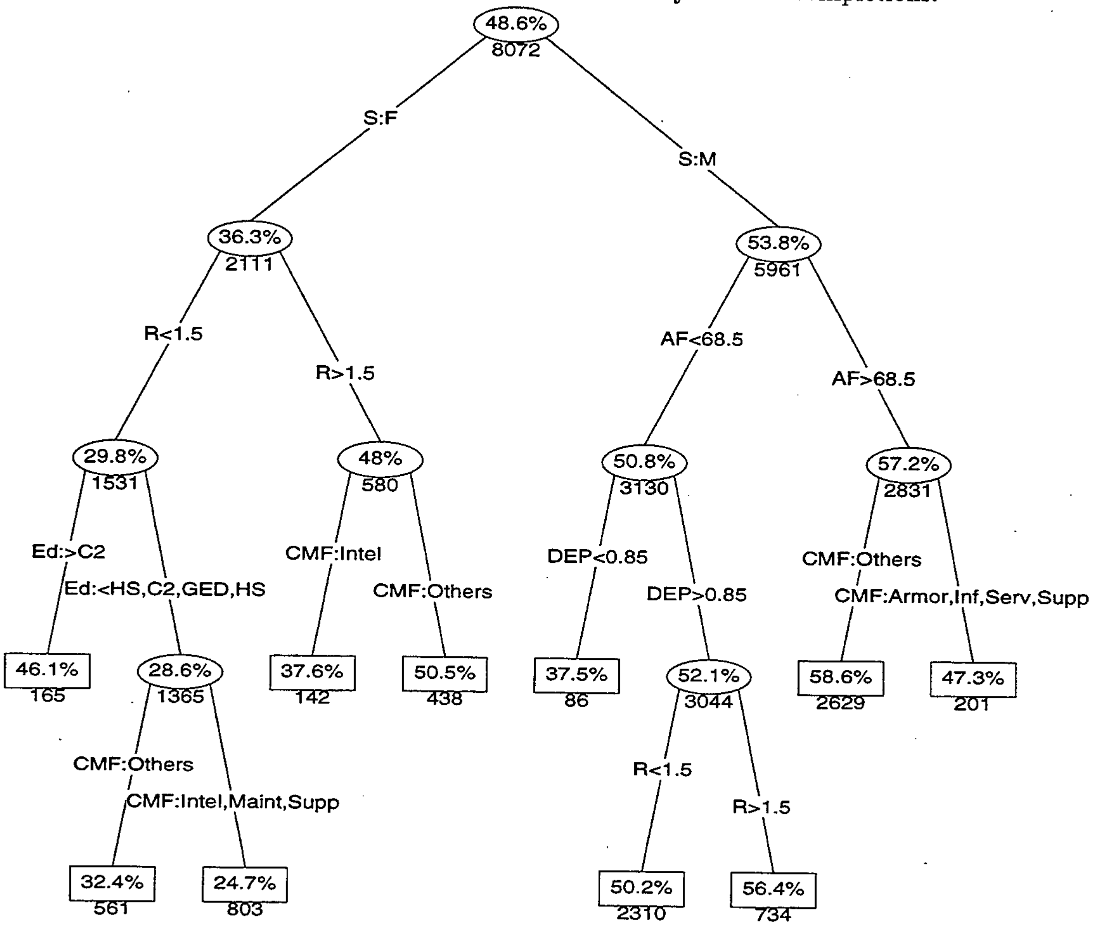
Table 1. Data available on completion of term

Calendar year	Term(s) missing	% of recuits with missing completion information
1991	None	0
1992	6 year	6.6
1993	5,6 year	12.2
1994	4,5,6 year	50.5
1995	3,4,5,6 year	95.2

Presumably the tendency for first-term enlistees to actually complete their contract term may change over time. This in turn would suggest that current and future behavior should be best modeled by using the most recent possible data. Accordingly, it was decided to primarily use the calendar year 1993 data in the CART model for completion of term. This data set is complete, except for the enlistees with five- and six-year terms, as noted in Table 1. To estimate their behavior, the six-year term enlistees from 1991 were pooled with the five-year term enlistees from 1992. CART was used to build a classification tree describing completion of term using this pooled data set. The variables used in building this tree were AF (AFQT score), R (race), S (sex), L (length of term) Age (age in months), DEP (time spent in the Delayed Entry Program), CMF (career management field), and Ed (education level). The resulting best ten-leaf tree is presented in Figure 3.

In total, 48.6% of these enlistees completed their first term (from the root node). This root splits first on sex, with females going left and males going right; the completion rate

Figure 3: CART predictions for five- and six-year term completions.



This tree is used to predict completions of five- and six-year terms in the 1993 data, for which the true completion status is not available. There was a 46.1% rate of term completion among white females with more than two years of college who entered in 1991 for six years or in 1992 for five years. Therefore we assumed that 46.1% of similar women who entered in 1993 for five or six years term would complete their terms.

for males (53.8%) is considerably larger than the rate for females (36.3%). These nodes are subsequently split, giving a total of five leaves for females and five for males. Both sexes split on race and on CMF (at different levels); females also split on education, while males have splits on AFQT score and DEP length. It is interesting that L, length of term, is not a splitting variable for either sex. The two different terms of service behave relatively similarly, for both sexes.

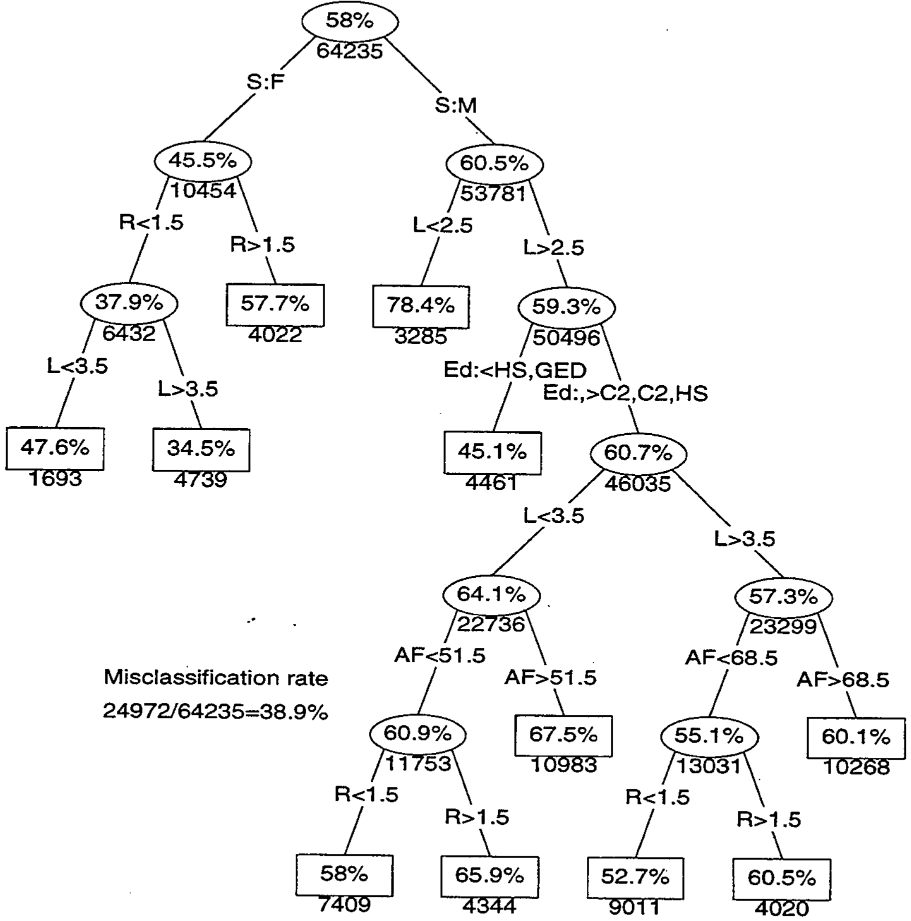
This tree in Figure 3 was used to predict completion of term for 1993 five- and six-year enlistees. That is, if we look at the lower left leaf we find that 46.1% of the white females with more than two years of college education completed their term. Thus, 46.1% of the 1993 cohort falling in this category, with five- or six-year terms, were randomly selected as completing their terms. Similarly, from the rightmost leaf we see that 47.3% of the males, with AFQT score above 68, and CMF of Armor, Infantry, Service or Support completed their term. For the 1993 five- and six-year term enlistees, 47.3% of those falling in this group were randomly selected as having completed their terms. Each of the other eight terminal leaves were used in the same manner.

This gives a 1993 data set with known completions of term for the two-, three-, and four-year term enlistees, and predictions of completions of term for the remainder, which will be referred to as the augmented 1993 cohort. This data set was then used to build a CART tree modeling completion of term; all the variables available were used. Cross-validation was then used on this tree several times and the results plotted (not shown). Recall that cross-validation *randomly* splits the data set into pieces for validation; thus redoing the cross-validation typically gives a different plot. Because the data set is large, though, these plots are extremely similar. The last major drop in deviance occurs with an eleven-leaf tree, so this is the size chosen for the CART model of completion of term. The best eleven-leaf tree is presented in Figure 4.

The root node contains the full data set; this group is then split on the sex variable (because this gives the largest drop in deviance). The larger collection of males has a higher proportion of persons completing their first term of service. The females then are further split on white race versus the others, with the white females further split on shorter terms of service (two or three years) versus the longer terms. Males also split on length of term (twice), education, AFQT score and finally on race. Three of the eleven leaves describe females, with the remaining eight describing males. The misclassification rate for this tree is 38.9%, as compared to 42% at the root.

For comparison, a tree describing the current classification groups (for the augmented 1993 cohort) is presented in Figure 5. This tree has a misclassification rate of 39.6%, incorrectly classifying 457 more individuals than the CART tree. Another way of comparing these two trees is to contrast their performances in predicting the completion of terms for

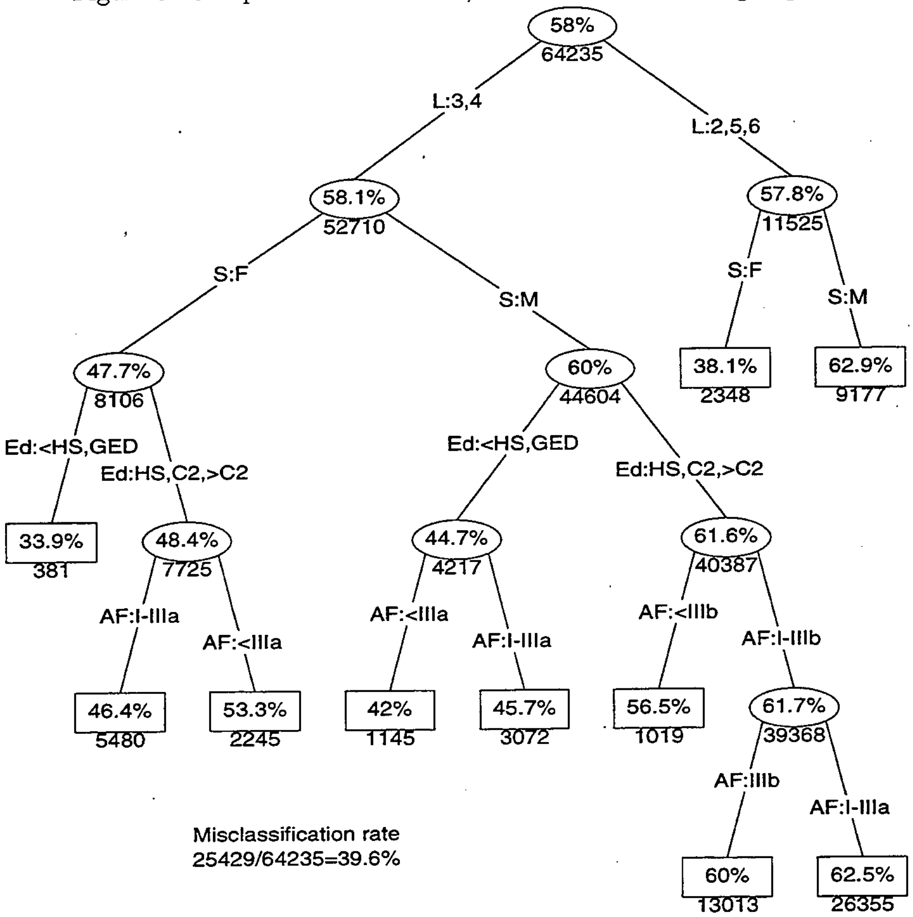
Figure 4: CART tree describing completion of term.



This tree gives the proportion of recruits completing their term in the 1993 (augmented) data. The smallest rate, 34.5%, was among while females with term lengths (L) greater than three; the largest rate, 78.4%, was among men with two-year terms. The misclassification rate is acheived by assigning to each recruit the predominant status in the leaf into which he or she falls.

later yearly cohorts (recall that completion of term is known for two- and three-year term enlistees in 1994, and for two-year term enlistees in 1995). Any given tree can be used to predict the numbers to complete their terms in each leaf (call this the expected number, e_i, for leaf i). Simple counting then can be employed to find the actual numbers to complete their terms in the same leaves (call this the observed number, o_i, for leaf i). Then the total number of errors in classification made by the tree is $\sum_i |e_i - o_i|$, the sum across the

Figure 5: Completion of term tree, current characteristic groups.



This is the analogous tree, using the current characteristic groups. Notice that the misclassification rate is slightly larger here than in the previous figure, and that the group sizes are less similar.

leaves of the magnitudes of the differences between the numbers expected and the numbers observed. This in turn can be divided by n, the number of enlistees classified, to give a total percentage error.

The 1994 cohort includes 27027 enlistees with two- or three-year terms. Using the measure just defined, the CART tree has a total percentage error of 0.8%, while the characteristic groups tree has a total percentage error of 4.8%. The 1995 cohort includes 2388 enlistees with two-year terms. The total total percentage error for the CART tree is

1.9% and is 17.4% for the characteristic groups tree. It is expected that basing forecasts on the CART tree groupings given in Figure 4 should perform better than forecasts based on the characteristic groups tree in Figure 5.

5 CART descriptions of re-enlistments

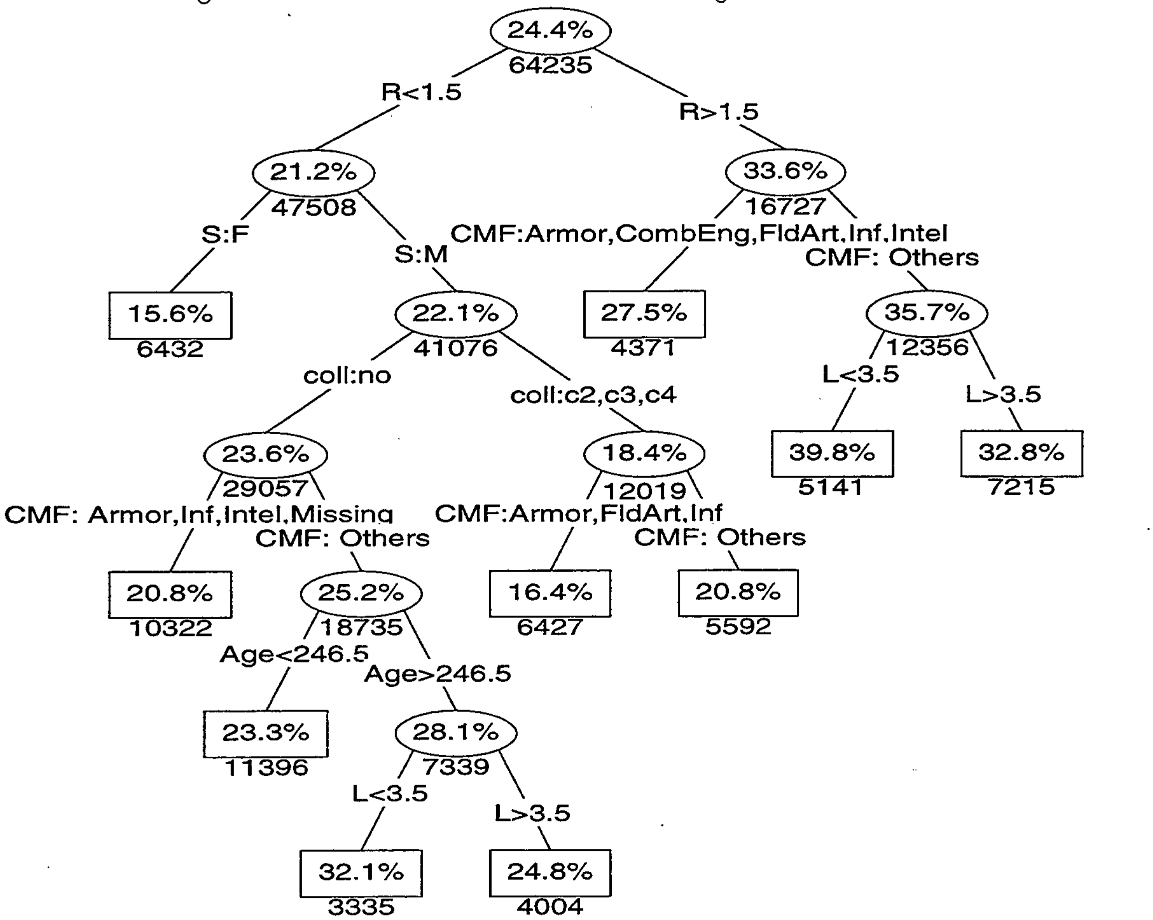
This section describes two different CART descriptions of re-enlistment of non-prior service enlistees. The first of these uses a binary variable (0 means no, 1 means yes) to represent the re-enlistment decision. Like the completion of term variable, this decision is not known with certainty for all term lengths for the 1992 and later cohorts. This problem was handled in the same way as described above for completion of term. The 1991 six-year and 1992 five-year term enlistees were used to build a CART description of re-enlistment; this tree was used in turn to predict the decisions (assigned randomly) of 1993 five- and six-year term enlistees, resulting in a new augmented 1993 data set.

The CART tree describing re-enlistment grew to 262 nodes, using the program defaults. The cross-validation plot (not shown) shows a relatively flat minimum for tree sizes from about 20 to 40 leaves, although most of the reduction in deviance is achieved with a tree size of 10. Accordingly, the best ten-leaf tree was grown and is presented in Figure 6. The first split is on the race variable, with the second splits given by sex and CMF. Interestingly, for white males, the college fund variable separates enlistees into those with no college fund (of whom 23.6% re-enlist) versus those who do take the college fund (of whom 18.4% re-enlist); the re-enlistment rate is probably lower for this latter group to allow them to take advantage of their college funding. The tree also indicates that those with shorter terms tend to re-enlist at a higher rate than those with longer terms.

CART was also used to describe a three-level variable involving re-enlistments of first term enlistees. Logically, each enlistee may

- Neither complete term nor re-enlist,
- Complete term but not re-enlist, or
- Complete term and re-enlist.

Figure 6: Best ten-leaf CART tree for modeling re-enlistments.



This tree is best (in the cross-validation sense) for modeling re-enlistments. It is similar to, but not identical to, figure 4.

It should not be possible to fail to complete the first term, yet still re-enlist. Because of the rules employed in interpreting the trailer records in the small tracking file, a small fraction of 1% of the full set of records actually indicated this behavior. These records were deleted for the CART modeling described below.

Again, the 1991 and 1992 experience for five- and six-year term enlistees was used to predict the behavior (on this three-level variable) for the 1993 cohort. This augmented data set was then used to describe this three-level variable. The default tree for this model produced 248 terminal leaves (using 64,138 enlistees from the augmented 1993 data set).

A cross-validation plot shows a quite broad minimum, spanning trees with about thirty to seventy leaves, although most of the reduction in deviance is accomplished with fourteen leaves. This was chosen as the most practical size and CART was used to build the best fourteen-leaf tree, presented in Figure 7.

Recall that the variable being modeled has three rather than two different possible outcomes. In any node, then, we have not only the total count of persons, but also how many of these fell into groups 1, 2 and 3 just defined. The plot of the tree now gives the total number of enlistees inside the node symbol itself, with three (rounded) decimal fractions presented below the node. In each case the ordering of these fractions is (group 1, group 2, group 3).

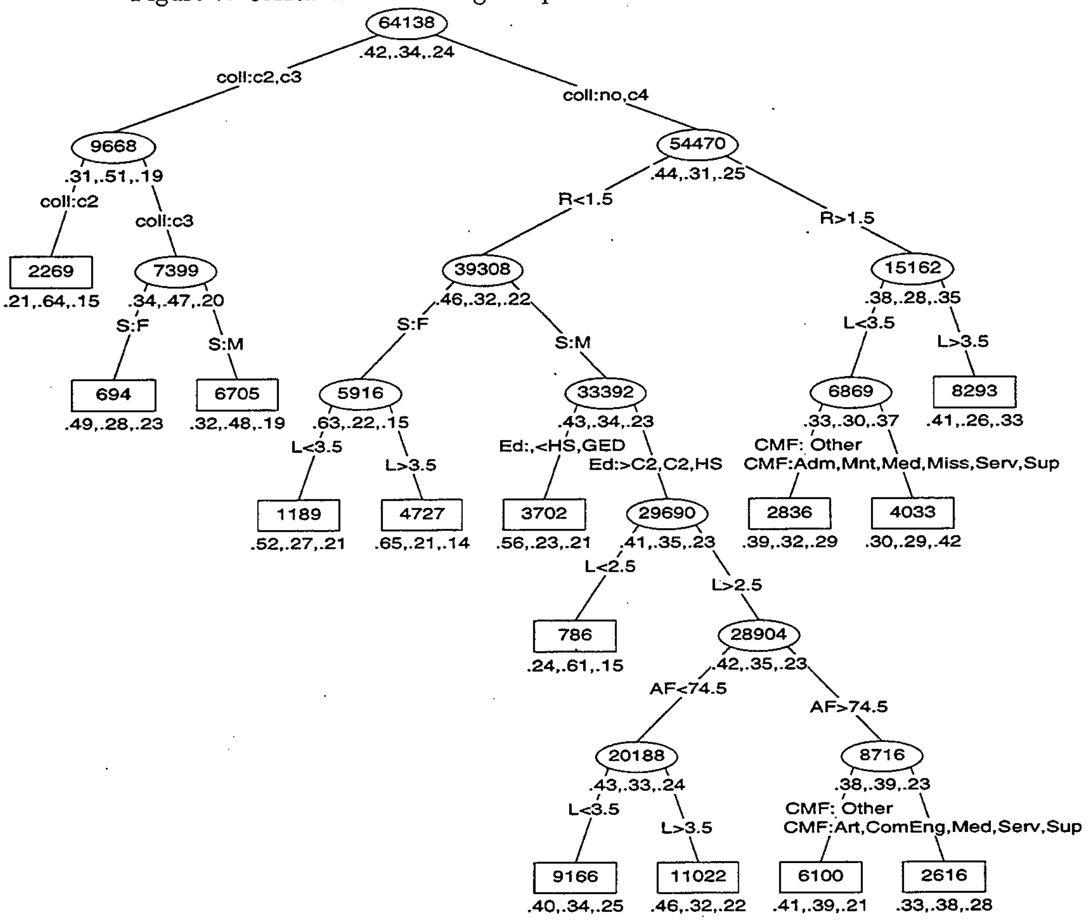
Thus, at the root in Figure 7 we can see that 64138 enlistees are included, of whom 42% neither completed their term nor re-enlisted, 34% completed the first term but did not re-enlist, while the remaining 24% both completed their first term and re-enlisted. Because of the rounding, the fractions below the nodes may not exactly total 1.00.

This rather large tree splits first on the college fund variable, with two- and three-year college fund recipients going left and four-year or no college fund going right. In the two- and three-year group (9668 enlistees), the majority (51%) complete their term, and only 19% re-enlist. This group is then further split into the two-year term enlistees (left) versus the three-year term enlistees. Almost two-thirds of the two-year term enlistees complete their term, while only a plurality (47%) of the three-year people do the same. The three-year group is then split on sex, with (again) males much more likely to complete the first term.

In the group including four-year or no college fund (54470 enlistees), most (44%) do not complete their first term, while 25% both complete the first term and re-enlist. This group is then split on race, with whites going left and all others going right. The whites are then split on sex, with the majority of white females (63% of 5916) not completing the first term, while only 43% of the white males fail to complete their first term.

The white females then split on length of term, again with the shorter terms having a smaller probability of not completing the first term. The white males subsequently split (twice) on length of term, AFQT score and CMF. The non-whites (four-year or no college fund) also split on length of term and CMF. They produce the leaf (two- or three-year term,

Figure 7: CART tree modeling completion of term and re-enlistments.



This tree shows recruits with three outcomes: attrition, completion without re-enlistment, and completion and re-enlistment. Beneath each node are the proportions in those three groups, in that order. For example, among those with the two-year college bonus (coll2; leftmost node), 21% underwent attrition, 64% completed their term without re-enlisting, and 15% completed their term and re-enlisted.

CMF of Administration, Maintenance, Medical, Service or Support, as well as Missing) which has the highest rate (42%) of both completing the first term and re-enlisting. These 4033 enlistees include 421 with MOS (and therefore CMF) missing.

6 Describing early-term attrition

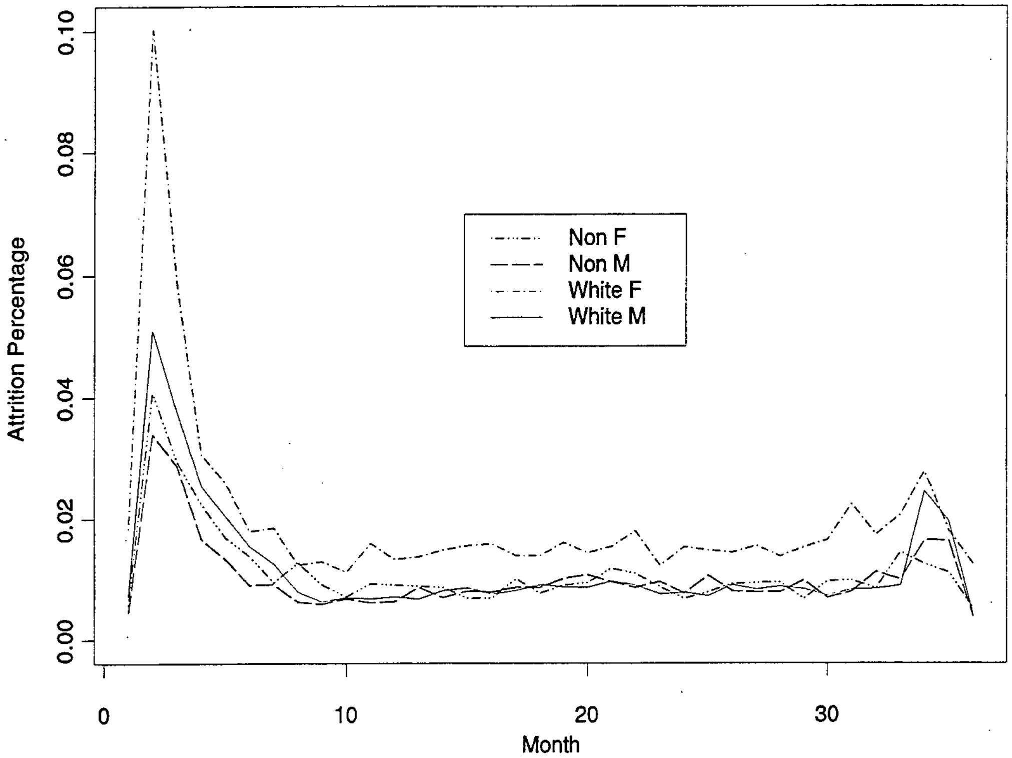
DCESPER and others have observed that a large fraction of first-term attrition takes place in the very earliest months of the term. It was asked whether there were differences in early-term attrition rates between different sub-groups (by age, sex and so on) of recruits. We can show these early-term attrition rates by means of *attrition profiles*, graphs of attrition rates by month. Figure 8 shows attrition profiles from the 1993 data (unmodified) for the four race and sex sub-groups. (Here, as elsewhere, we have divided race into "white" and "non-white.") On the horizontal axis is the month of the term; on the vertical is the proportion of those who reached that month who dropped out during that month. (So, for example, "3%" in month eight indicates that 3% of those who completed month seven dropped out in month eight.) We have chosen race and sex because these turn out to be the best predictors of attrition in the early part of the term.

The things to notice about this graph are, one, that every group shows a sharp peak of attrition in the first few months; two, that in each of the groups the attrition rate then settles down to a level that is more or less constant after about month nine; and three, that white females have both a higher peak and a higher "steady-state" level than the other three groups, which are close to indistinguishable.

These observations are supported by a number of statistics. The proportions of recruits leaving within the first nine months are 12.6% for non-white males, 16.1% for non-white females, 18.0% for white males and 27.8% for white females. It is interesting to observe that non-white females have a lower attrition percentage than white males in the first nine months. Over 36 months, attrition proportions (this, of course, excludes two-year term recruits who finish their terms) for the same groups are, respectively, 35.7%, 42.3%, 40.7% and 62.1%. In these two cases, and elsewhere, we see that non-white males have the lowest rate, that white males and non-white females have roughly equal rates, and that white females have the highest attrition rate.

The "steady-state" attrition rate — that is, the rate that the group "settles down" to after the peak — is seen to be similar for the first three groups. Among 1993 recruits completing their tenth month, the proportions that eventually left prematurely were 27.1% for non-white males, 32.7% for non-white females, 29.0% for white males and 50.0% for

Figure 8: Attrition Profiles, 1993 data, by race and sex.



This picture shows the rates of attrition by month, among males and females and whites and non-whites. Note the early peak in all groups, and the higher steady-state rate among white females.

white females. These rates do appear to be fairly steady. There is no evidence of an increasing trend for the first three groups (based on linear regression); there is some slight evidence, perhaps, that the steady-state rate for white females may increase somewhat over time.

7 CART descriptions of nine-month behavior

We are now in a position to construct CART trees of nine-month behavior. (We choose nine months simply by examining figure 8; a choice of eight of ten would not affect our conclusions materially.) In this case the two possible outcomes are survival through nine months, and attrition in or before the ninth month. The predictor variables, of course, are as before.

Figure 9 shows the best ten-leaf classification tree from the 1993 data. Interestingly, in this tree there is only one leaf for all non-white recruits. Two leaves describe the white females and the remaining seven describe white males. Partly this is a consequence of stopping at ten leaves (in the twelve-leaf tree, the non-white males and non-white females are split). In this tree we see, as we expect, high attrition among white females, low attrition among non-whites, and medium values of attrition among white males.

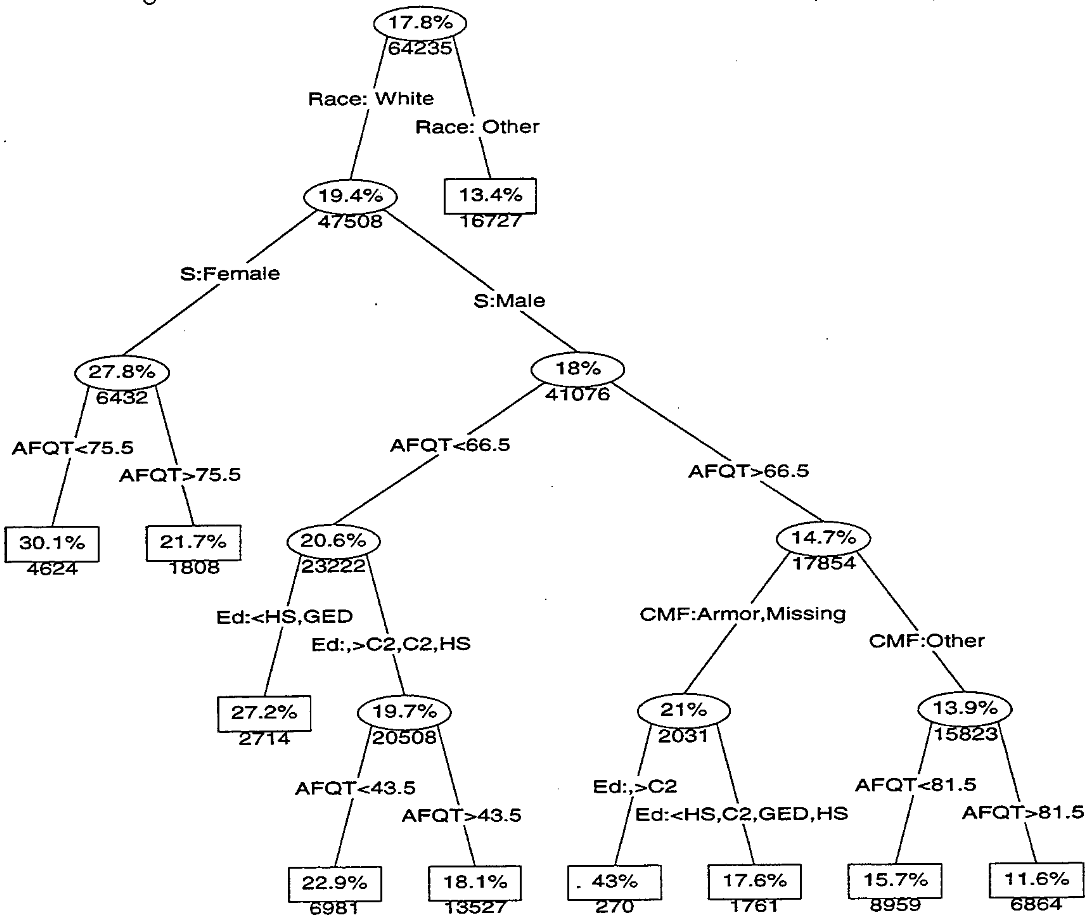
Another tree worth looking at is one where the race and sex variables are replaced by a single variable taking on the four age and sex combinations. In this tree (figure 10) the biggest decrease in deviance is achieved by first splitting off the white females from the other recruits. As before, these females have a high attrition rate. The second steps splits off the low-attrition non-white males, and the remaining parts of the tree describe the behavior of the white males and non-white females.

8 Other Differential Factors

Some of the difference in female attrition rates may be due to differential enlistment rates. In particular, the average length of term at sign-up is shortest for non-white males (3.52 years in 1993); in the middle for white males (3.58) and non-white females (3.91); and longest for white females (an average of 4.06 years). Some of the gender difference can be explained by differences in CMFs, since these tend to have differing term-lengths and some CMFs are prohibited to women. However, the racial difference is a mystery to us.

More than half of the men (54%) who joined the Army in 1993 signed up for terms of two of three years, but only about a quarter (28%) of women did so. Similarly, 24% of white females, but only 8% of non-white males, enrolled for terms of five or more years.

Figure 9: Best ten-leaf CART tree for nine-month attrition (1993 data).

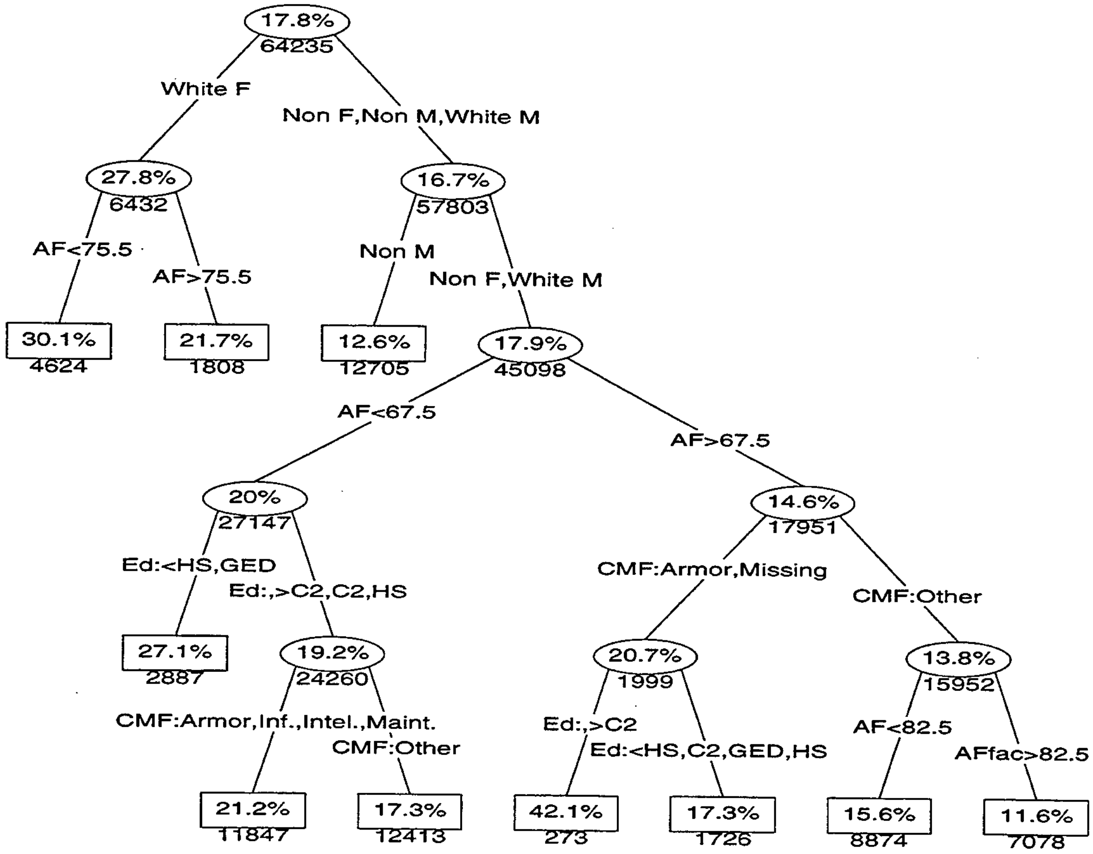


Still, there is a real gender and race differential, even when differing term lengths and CMF distributions are taken into account. Among those signing up for three-year terms, the proportions completing are 66.3% for non-white males, 66.1% for non-white females, and 60.5% for white males, but only 45.6% for white females. (Among those with four-year terms, the corresponding percentages are 61.1%, 55.8%, 56.0%, and 36.6%.)

9 Conclusions and Recommendations

The current characteristic groups can be improved upon. We have suggested a new set of groups that reduce misclassification rate to a small extent and that should therefore be

Figure 10: Ten-leaf tree for nine-month attrition, race and sex categories



useful to the Army's stength management system. The bulk of the improvement probably comes from the inclusion of race information. Non-whites have lower attrition, higher term completion and higher re-enlistment rates than whites. Males have lower attrition, and higher completion and re-enlistemnt rates than females; part of this may be due to the increased propensity of women to sign up for long terms. While the best characteristic groups for predicting re-enlistment or nine-month are slightly different from those for predicting term completion, the race and sex pattern remains. The college bonus programs appear to have little influence on term completion or early-term behavior, though they do affect re-enlistment. AFQT scores, education level, and Career Management Field all serve predictive roles as well. We recommend that the Army examine why women fare less well than men on average, perhaps by examining the reasons for attrition where those are known. Since there is a peak in attrition for all groups in the early months, those months would seem to be a good place for intervention strategies, intended to reduce attrition, to be aimed.

DISTRIBUTION LIST

1. Research Office (Code 09) 1
   Naval Postgraduate School
   Monterey, CA 93943-5000

2. Dudley Knox Library (Code 013) 2
   Naval Postgraduate School
   Monterey, CA 93943-5002

3. Defense Technical Information Center 2
   8725 John J. Kingman Rd., STE 0944
   Ft. Belvoir, VA 22060-6218

4. Therese Bilodeau (Editorial Assistant) 1
   Dept of Operations Research
   Naval Postgraduate School
   Monterey, CA 93943-5000

5. Prof. Samuel E. Buttrey (Code OR/Sb) 2
   Dept of Operations Research
   Naval Postgraduate School
   Monterey, CA 93943-5000

6. Prof. Harold J. Larson (Code OR/Jc) 2
   Dept of Operations Research
   Naval Postgraduate School
   Monterey, CA 93943-5000